The Central Limit Theorem

BIOS 6611

CU Anschutz

Week 3

What is the Central Limit Theorem?

Introducing the Central Limit Theorem (CLT)

- One of the most *fundamental* and *profound* concepts in statistics. Maybe even all of mathematics!
- In a nutshell: the sampling distribution of the mean of ANY DISTRIBUTION converges to a normal distribution, under certain (common) conditions

Why does it matter?

Why is the CLT important?

- In real world, there are all sorts of random processes. Often we do not know the underlying distribution, or it might be skewed.
- CLT says, with enough samples, can treat sampling distribution of the mean as normal, no matter the underlying distributions (if samples are IID)
  - IID = independent, identically distributed (come from same distribution)
- Can apply statistical methods that require normality assumption to any data with "large" sample size
- The CLT is partly why the normal distribution shows up so much in statistics

CLT Statement, Mathematical Formulas

CLT Statement

Mathematically, the CLT states if $(X_1, \ldots, X_n)$ are IID, $E[X_i] = \mu$ exists, and $Var[X_i] = \sigma^2$ exists, then:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \rightarrow_d N(0, 1)$$

(d = "converges in distribution")

Another way to look at CLT Statement

Another way to think of the CLT is

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \tag{1}$$

for n large (rule of thumb: $n \geq 30$, but can be much larger)

Points of Clarification

Point of Clarification

The CLT is a statement about the *sampling distribution of the sample mean* ($\bar{X}$), NOT the distribution of the sample (X_i).

The distribution of the *sample* gets closer to the population distribution as n increases.

Another point of clarification

If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ no matter the sample size.

If X_i are not normal (i.e., any other distribution with a defined mean and variance), then $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$ only for large n.

CLT with Simulations

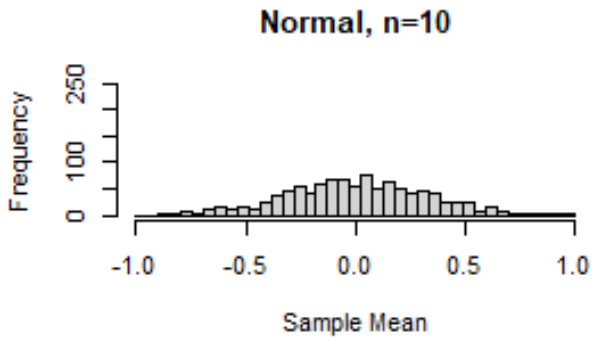
CLT with Simulations

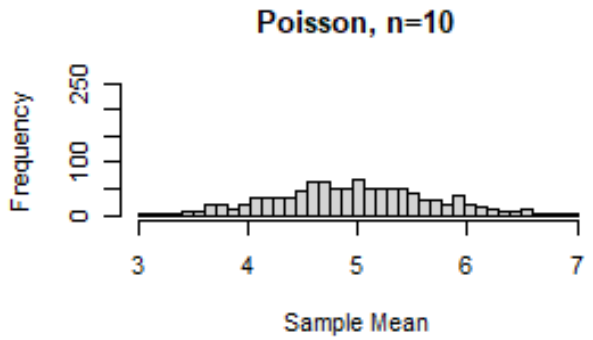
Let's show ourselves the CLT is true using simulations. We will compare the following scenarios across 1000 simulated data sets for each:

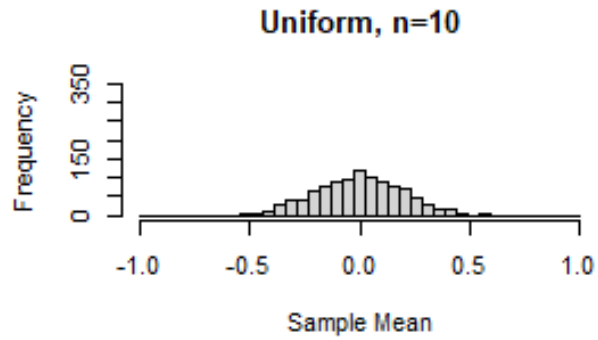
- Sample sizes: $n = 10, 30, 100$
- Distributions: Normal($\mu = 0, \sigma^2 = 1$), Poisson($\lambda = 5$), Uniform(-1,1)

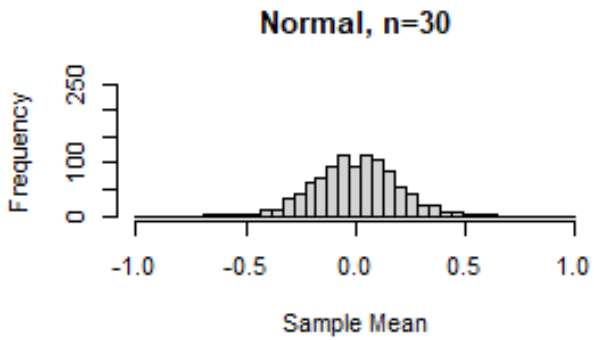
```
set.seed(6611) # set seed for reproducibility
mat_res <- matrix( nrow=1000, ncol=9 ) # create matrix to store results
nvec <- c(10,30,100) # vector of sample sizes to explore

for(i in 1:1000){
  # Simulate 100 observations (will take first 10/30/100 for estimates)
  normal_dat <- rnorm(n=100, mean=0, sd=1)
  poisson_dat <- rpois(n=100, lambda=5)
  uniform_dat <- runif(n=100, min=-1, max=1)

  # Store estimated means
  mat_res[i,1:3] <- sapply(nvec, function(x) mean(normal_dat[1:x]))
  mat_res[i,4:6] <- sapply(nvec, function(x) mean(poisson_dat[1:x]))
  mat_res[i,7:9] <- sapply(nvec, function(x) mean(uniform_dat[1:x]))
}
```

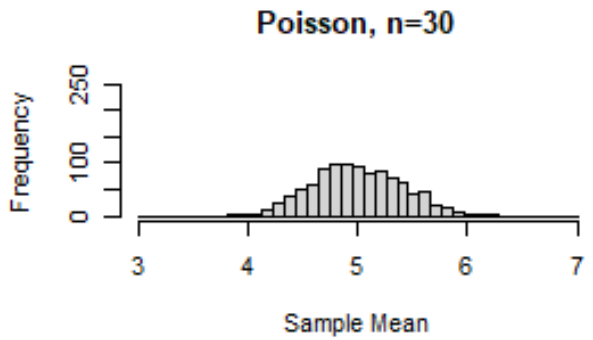
CLT with Simulations

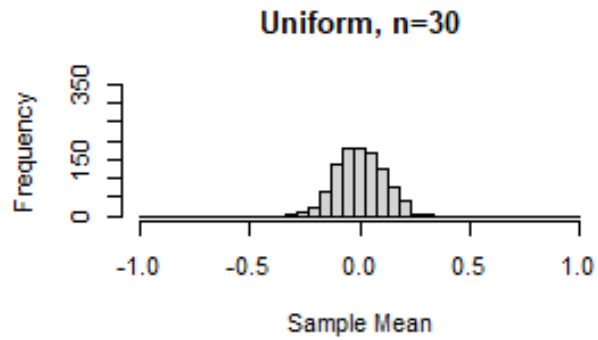


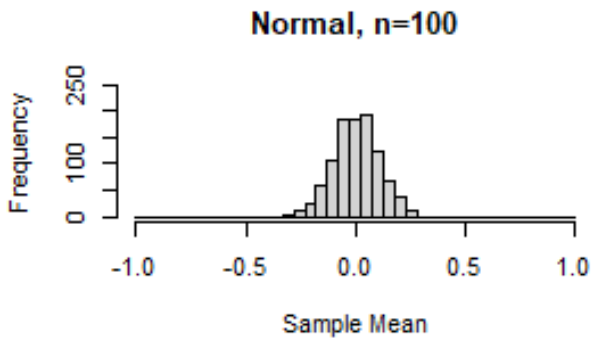


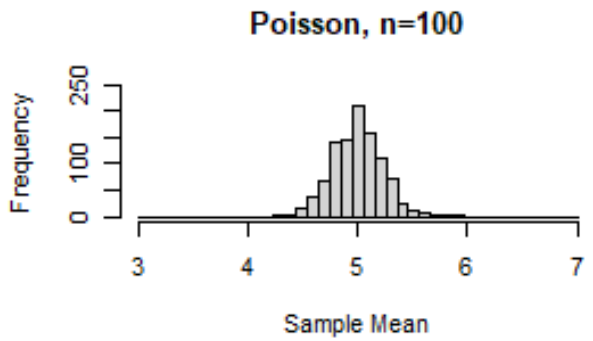


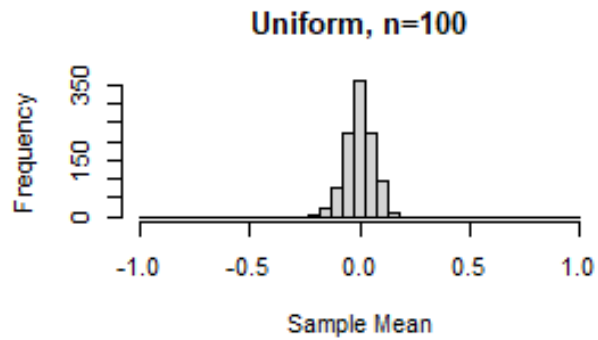












Proving CLT

You can prove the CLT mathematically several ways, however all of the proofs I have seen are non-trivial, so won't go through it in this class.

Involve applying Taylor's expansion and taking the limit:

$$f(x) = f'(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \ldots \tag{2}$$

Summary

- CLT is one of the most important results in statistics.
- Says that distribution of the sample mean is approximately normal for large sample sizes.
- Allows us to apply techniques developed for inference on normally distributed populations to any IID population with large enough sample size. This will include many of the tests we will encounter this semester!